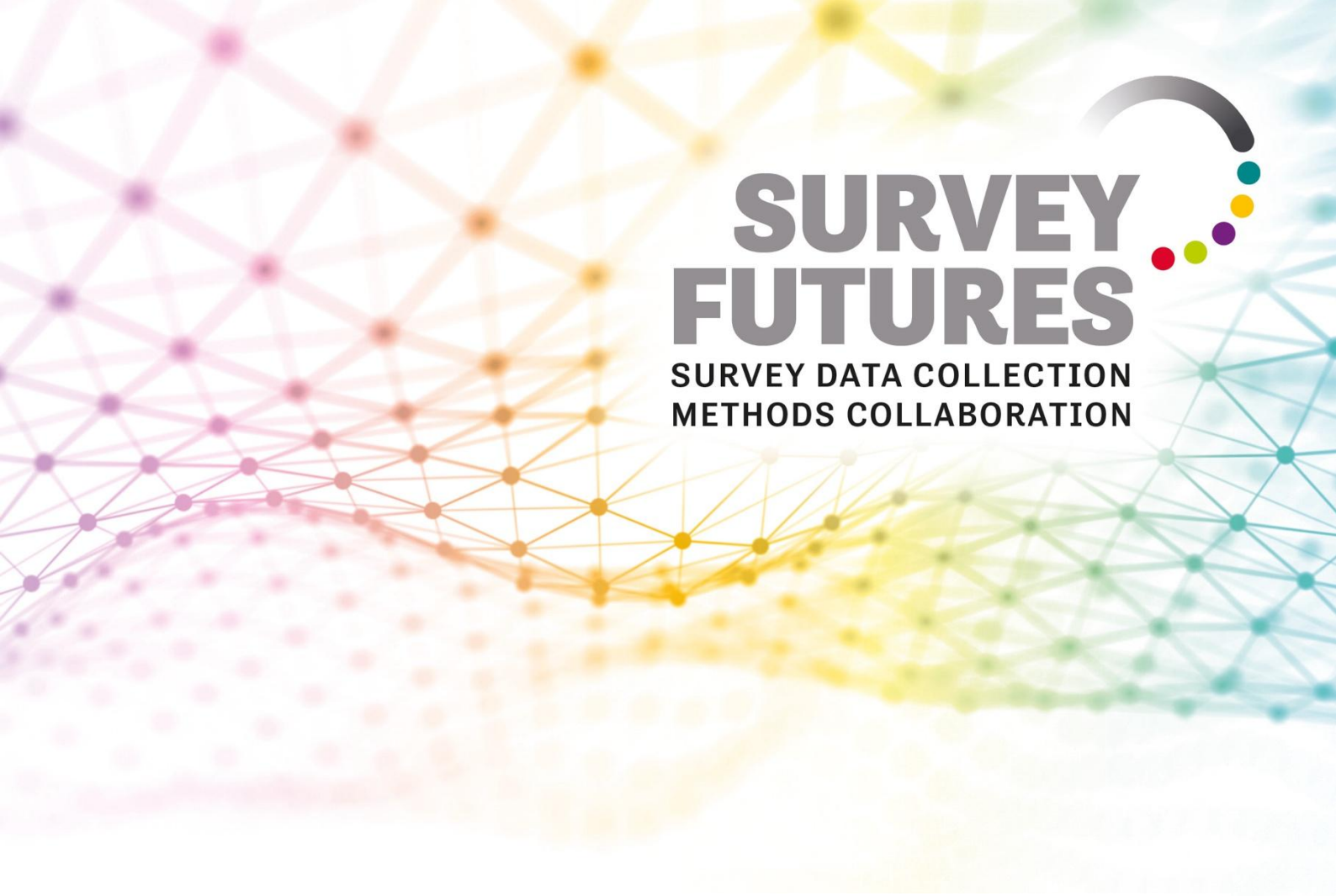


Working Paper 5:

Navigating Challenges in Occupation Data Collection in a Mixed-Mode Longitudinal Survey: Insights into the Look-Up Approach

Sebastian Kocar[1], Darina Peycheva[2], Matt Brown[2], Joseph W. Sakshaug[3], Claire Bhaumik[4], Lisa Calderwood[2]

[1]University of Queensland; [2]University College London; [3] Institute for Employment Research (IAB) & LMU-Munich, [4]Ipsos

June 2025

Acknowledgements

Survey Futures is an Economic and Social Research Council (ESRC)-funded initiative (grant grant ES/X014150/1) aimed at bringing about a step change in survey research to ensure that high quality social survey research can continue in the UK. The initiative brings together social survey researchers, methodologists, commissioners and other stakeholders from across academia, government, private and not-for-profit sectors. Activities include an extensive programme of research, a training and capacity-building (TCB) stream, and dissemination and promotion of good practice. The research programme aims to assess the quality implications of the most important design choices relevant to future UK surveys, with a focus on inclusivity and representativeness, while the TCB stream aims to provide understanding of capacity and skills needs in the survey sector (both interviewers and research professionals), to identify promising ways to improve both, and to take steps towards making those improvements. Survey Futures is directed by Professor Peter Lynn, University of Essex, and is a collaboration of twelve organisations, benefitting from additional support from the Office for National Statistics and the ESRC National Centre for Research Methods. Further information can be found at www.surveyfutures.net.

The research reported here forms part of Research Strand 5 of *Survey Futures*, led by Professor Lisa Calderwood, University College London, which focuses on complex measurement in self-completion surveys.

The Next Steps Age 32 Survey was designed and managed by the UCL Centre for Longitudinal Studies, supported by the Economic and Social Research Council [grant number ES/W013142/1]. Fieldwork was conducted by Ipsos.

Citation

Prior to citing this paper, please check whether a final version has been published in a journal. If so, please cite that version. In the meanwhile, the suggested form of citation for this working paper is:

Kocar S., Peycheva D., Brown M., Sakshaug J. W., Bhaumik C. & Calderwood L. (*2025*) 'Navigating challenges in occupation data collection in a mixed-mode longitudinal survey: insights into the look-up approach', *Survey Futures Working Paper* no. 5. Colchester, UK: University of Essex. Available at https://surveyfutures.net/working-papers/.

Contents

Navigating Challenges in Occupation Data Collection in a Mixed-Mode Longitudinal Survey: Insights into the Look-Up Approach

Abstract

Occupation data have traditionally been collected through interviewer administration using open-ended questions and manual office coding, with alternative approaches being developed in recent years. These include the look-up self-coding approach, which presents a range of new challenges that require further methodological exploration. This study investigates the feasibility and quality of the look-up approach by comparing it with traditional office coding in the 9th Sweep of the Next Steps longitudinal study, a mixed-mode survey. To assess the quality of the look-up occupation data, the study incorporated an experiment in which participants were asked to self-code their occupation but also to provide an open-ended description of their job, which was then manually coded by two independent office coders. We used two indicators of feasibility and data quality, namely the look-up coding rate and the agreement between the look-up and office coding, with look-up input metrics and demographic predictors used to identify potential methodological solutions. The results show that the look-up coding rates were higher in interviewer-administered modes (90%) than in the web mode (82%), with high office coding rates (99%) across all modes. Also, the agreement rate between look-up and office coding was significantly lower than between two office coders, which we critically assessed. Additional investigation showed that coding and agreement rates could be linked to look-up input metrics including lengthy job description keywords and 1-digit occupation code, as well as not entering job information (coding rates) and how well the respondents believed the look-up code described their job (agreement between the look-up and office coding). Importantly, the look-up input metrics largely explained the differences in coding rates between the modes. Based on the presented evidence, we propose that the optimal solution may be to supplement the look-up with office coding for respondents with missing or potentially less reliable look-up codes.

Keywords: occupation coding, look-up coding, coding agreement, coding rates, mixed-mode survey, longitudinal study

1. Introduction

Occupation is a key measure in many social surveys (Tijdens 2022). It serves as an important indicator of one's socio-economic status and has a significant impact on various aspects of life, including income, health, and lifestyle. It is typically measured by using a series of open-ended questions with the answers coded to a standard code-frame, for example, the Standardized Occupation Classification (U.S. Bureau of Labor Statistics n.d.), the UK's Standard Occupational Classification (SOC) 2020 (Office for National Statistics n.d.), or International Standard Classification of Occupations (ISCO [International Labour Organization 2010]). Occupation coding has been an important subject of methodological research which has sought to develop approaches for valid and reliable measurement of occupation in surveys (Hoffmann and Thomas 1995). Traditionally, occupational measures have been collected through interviewer-administered surveys with responses manually coded by professional coders (e.g., Lyberg and Dean 1992). However, as online surveys become increasingly common, respondents are now often asked to provide occupational information without the assistance of an interviewer, which may have consequences for the quality of the collected data (Conrad et al. 2016).

Collecting and coding occupation data presents several challenges, especially in self-completion studies such as web surveys (e.g., Peycheva et al. 2021). For example, occupations can be as diverse as survey respondents and different individuals may describe the same occupation in different ways (Simson et al. 2023). Occupation coding is typically conducted post-survey without the possibility of further probing (Simson et al. 2023) and respondents can provide unspecific or invalid answers that cannot be coded (Belloni et al. 2016; Conrad et al. 2016). It is also difficult to achieve high accuracy with fully automated coding (Gweon et al. 2017). As a result, the agreement between two coding experts, or between a professional coder and an automated coding algorithm, can be relatively low (Russ et al. 2023[1]). Also, as online surveys become increasingly common, respondents are now often asked to provide occupational information without the assistance of an interviewer, which may have consequences for the quality of the collected data (Conrad et al. 2016).

Respondent self-coding[2] during questionnaire completion has been developed as an alternative to the collection of open-ended descriptions of occupation for subsequent office coding, which is still considered the "gold standard" approach (Burstyn et al. 2014). It aims to address the challenges associated with the accuracy of occupation coding, discussed above, and the cost- and time-efficiency of data collection and coding (e.g., Schierholz et al. 2018). A so-called *look-up method* (see Tijdens 2015) can be used in both self-completion and

[1] In their study, the agreement at the 3-digit level between one coding expert and an automated coding algorithm was about 60%, and the number of codes at that level in the U.S. Standardized Occupation Classification is 97 (Russ et al. 2023).

[2] In addition to (interactive) self-coding, terms such as self-identification or self-classification are also used in the literature (Mannetje and Kromhout 2003; Tijdens 2016), although the meanings slightly differ.

interviewer-administered surveys. It typically includes the following steps: (1) the respondent enters information about their job, such as job title and description of duties, (2) a search function algorithm uses the information to generate a list of the most likely occupations, and (3) the respondent selects the most suitable category (Simson et al. 2023)[3]. In the case of interviewer administration, the interviewer enters the information and assists respondents with assigning the best category (Peycheva et al. 2021).

While existing research on occupation coding using the look-up in different survey modes has provided some valuable insights, several research gaps remain to be investigated. We investigate such gaps by addressing the following research questions (RQs) regarding the use of the look-up function in a mixed-mode survey[4], with a primary focus on self-completion without the assistance of an interviewer. In this study, a look-up approach was used, but in addition, participants were also asked to provide open text descriptions of their job, which were manually coded by two independent expert coders.

> ***RQ1:*** *What proportion of respondents successfully select an occupation code using the look-up method (i.e., the look-up coding rate) and how accurately do they feel this code describes their occupation? What proportion of open-ended job descriptions are successfully office coded? Are there any differences in coding rates and the respondent's assessment of the accuracy of the code selected between self-administration (Web) and interviewer administration (F2F, CATI)? What are factors that affect look-up coding rates?*

We will first provide an update on the past evidence regarding coding rates presented in similar research (cf. Hacking et al. 2006; Peycheva et al. 2021; Schierholz et al. 2018). The coding rate serves as our primary indicator of the feasibility of the look-up approach, and we will compare the rates between the survey modes. We will contrast this to the proportion of occupations where open-text descriptions were manually coded. Additionally, we will explore both procedural and respondent factors that affect look-up coding rates.

> ***RQ2:*** *What is the agreement between the look-up and office codes for respondents who were assigned both, and how does it compare to agreement rates between two office coders? Are there any differences between the modes (Web, F2F, CATI)? What are the factors that affect the agreement between look-up and office coding, as well as between two office coders?*

---

[3] In some instance, respondents are able to repeat the search process to regenerate the list if none of the initially proposed categories is considered suitable.

[4] The following modes have been used to collect occupation data: Web (CAWI), Computer Assisted Personal Interviewing (CAPI), Live Video Interviewing (LVI), and Computer Assisted Telephone Interviewing (CATI). Considering that CAPI and LVI can both be classified as face-to-face (F2F) modes, they will be treated as one category, especially since the number of LVI interviews was small (n=8). To discuss the differences between self-completion and interviewer administration, F2F and CATI can be combined into interviewer-administered modes, particularly given the relatively small number of CATI interviews (n=114). Notably, the design of this study was not experimental and so respondents were not randomly assigned to modes.

We aim to evaluate the quality of the look-up coding by comparing the look-up codes to those assigned by expert coders at four SOC levels. Since all open-descriptions were double-coded by office coders, we can compare the agreement between look-up and office coding and the agreement between two office coders as our second indicator of coding quality[5]. This methodology mirrors that of Schierholz et al. (2018), who compared telephone interview look-up coding and office coding[6]. Additionally, we examine the factors which affect the level of agreement including procedural aspects, respondent characteristics, and the respondent assessments of the accuracy of the selected look-up code. This evidence can help identify scenarios where collecting additional open-ended job descriptions could enhance the overall coding accuracy.

2. Literature review

2.1 Emergence and use of the look-up approach for occupation coding

Occupation data has traditionally been collected in interviewer-administered modes (typically F2F) using open-ended questions, and the answers then manually coded by expert office coders post-survey (Lyberg and Dean 1992). However, new approaches have since emerged, driven by new technological and methodological developments. Alongside the shift towards self-completion surveys there have been other developments in collecting occupational information and occupation coding, including software-assisted office coding, self-coding during the interview, and fully-automated coding post-survey (e.g., Hacking et al. 2006; Mannetje and Kromhout 2003; Ossiander and Milham 2006; Safikhani et al. 2023; Schierholz and Schonlau 2021).

Self-coding, which includes using a look-up search function, is a relatively newer approach (e.g., Brugiavini et al. 2017) that has been used in interviewer-administered (e.g., Schierholz et al. 2018), self-completion, and mixed-mode surveys (e.g., Peycheva et al. 2021), and applied in both cross-sectional (e.g., Hacking et al. 2006) and longitudinal contexts (e.g., Peycheva et al. 2021). The look-up approach can be used as a stand-alone coding method (e.g., Tijdens 2016) or combined with collection of open-descriptions of jobs and subsequent office coding. In the case of using both approaches, office coding can be used as a supplement to assign codes for respondents who could not self-select an occupation code from the generated list (see Peycheva et al. 2021). This strategy is similar to how automated coding may be combined with manual office coding. Implementing both coding approaches together in the same study can also result in higher overall coding rates, greater coding accuracy, assessment of coding reliability, and it may enable further evaluation and development of either standalone approach (Burstyn et al. 2014).

[5] This second indicator is based on the premise that office coding of open-text descriptions of jobs has traditionally been considered the most accurate approach of occupation coding.

[6] Similarly, Russ et al. (2023) assessed the feasibility of fully automated coding by comparing (i) agreement between two coders and (ii) one coder and an automated coding algorithm.

2.2 Procedural aspects of the look-up approach

An open question related to the look-up approach is how its characteristics and procedures affect the quality of the collected data. The approach typically relies on a search function that uses algorithms, such as machine learning and other techniques, to process provided job information and generate a list of the most likely codes from code-frames, such as the Standardized Occupation Classification (U.S. Bureau of Labor Statistics n.d.). In the past, self-classification of occupation was considered a highly challenging approach due to respondents' inability to reliably identify the most suitable category for their occupation (Mannetje and Kromhout 2003). This challenge has become less significant since the introduction of different algorithms and advanced computerization, which assists respondents in identifying the most relevant category, but many challenges remain.

One of the main issues that questions with a long list of answer categories[7] face is presenting the list of categories in a way that does not overburden respondents. Herzing (2020) identified a text box combined with a drop-down box as having the most positive impact on response burden and measurement. While search trees have been tested in the occupation coding space, they have been identified as relatively cognitively demanding (Tijdens 2014). For that reason, the format that combines a search function and an automatically generated list of most likely categories has been used for collecting occupation data using the look-up approach in most studies to date (Hacking et al. 2006; Peycheva et al. 2021; Schierholz et al. 2018; Simson et al. 2023; Tijdens 2016). However, this methodology does not come without its challenges – in addition to some cognitive demand, it is also time-consuming for respondents, and the choice set can be incomplete to the extent that respondents cannot identify the code for their occupation (Tijdens 2016). For that reason, some studies provide instructions to update the search to either narrow down the list and/or generate a more relevant list (Peycheva et al. 2021).

The number and content of questions typically asked of respondents for occupation coding purposes can be quite diverse (see Tijdens 2014, p. 12, for a detailed review). However, in the case of the look-up approach, there appears to be more consistency in their content. In practice, answers to questions such as the respondent's job title, information about their occupational activity, descriptions of what respondents mainly do in their jobs, and requests for similar information are typically used (Peycheva et al. 2021; Schierholz et al. 2018; Simson et al. 2023; Tijdens 2016). The information provided by respondents is then used by algorithms to generate a list of possible occupations (Simson et al. 2023), and different studies have reported varying numbers of occupational codes that are proposed to the respondent

[7] For example, the highest SOC level can include between 413 (4-digit, the UK) and 840 categories (6-digit, the U.S.).

by the algorithm[8] (see Peycheva et al. 2021; Schierholz et al. 2018;). Since the appropriateness of the proposed list of occupations depends on the respondents' answers about their jobs, existing research has also investigated the effect of the length of answers on coding reliability. Most studies employing different occupation coding approaches have quite counterintuitively found that longer job descriptions, measured in terms of the number of characters or words, are equally or less reliably coded than shorter descriptions (Conrad et al. 2016; Helppie-Mcfall and Sonnega 2018; Massing et al. 2019). Lastly, the amount of time spent using the look-up function can play a role in whether an occupation code is selected, with those who could not assign an occupation taking more time (Peycheva et al. 2021).

2.3 Relative quality of data collected with the look-up method

There are three main indicators of occupation data quality described in the literature, which can also be used to directly compare different coding approaches. First, either inter-coder reliability or agreement rate between two coders (or coding approaches) have been traditionally used when coding open-ended answers to occupation questions (see Kim et al. 2020). Second, the coding rate as a proportion of all respondents in work with an assigned occupation code, is another indicator (see Gweon et al. 2017; Schierholz et al. 2018). And third, in the case of automated coding, coding accuracy is typically assessed by exploring agreement rates between algorithms and manual coding (Gweon et al. 2017). These indicators of occupation coding quality can also be used to assess look-up coding, as we show in the forthcoming analyses.

The existing literature reports both differences in the quality of occupation data collected with the look-up approach (e.g., coding rates), as well as differences between modes of survey administration. Higher coding rates have been reported for office coding, considered the "gold standard" approach, compared to coding during the interview (Burstyn et al. 2014). For instance, Peycheva et al. (2021) reported that open descriptions of occupations could be coded for 99% of all respondents using office coding, including more than 90% of free-text answers from respondents/interviewers who could not select a suitable occupation from the look-up (representing approximately 18% of the whole sample in their survey). Other studies have reported similar results (e.g., Schierholz et al. 2018).

One important aspect that has been addressed in the automated coding literature (e.g., Russ et al. 2023) but not extensively in self-coding literature is comparing agreement (i) between two office coders, and (ii) between one office coder and a non-office coding approach (i.e., self-classification using a look-up function). In the self-coding space, Schierholz et al. (2018) compared the agreement between two professional coders (who used open-descriptions) and between coding conducted by a telephone interviewer (who used a look-up method) and

[8] This number is determined by the survey designer. For example, in Next Steps, all SOC codes with a certain closeness score were displayed, with a cap of 35 results.

by a professional coder (who used open-descriptions). They reported very minor differences in agreement, indicating that using occupation coding solutions like the look-up function is promising.

3. Methods

3.1 Data

The data used in this study are from the 9th Sweep of Next Steps, a longitudinal study in England which tracks the lives of approximately 16,000 participants born in 1989-90. Participants were initially recruited in 2004 when they were aged 13-14. They were surveyed annually until 2010, with data collected from them again in 2015-16 (Sweep 8) and 2022-23 (Sweep 9). Fieldwork for Sweep 9 was conducted by Ipsos. The study collects information about a broad range of topics including family life, health and wellbeing, education, social participation and attitudes. Collecting information about employment, including occupation[9], has also been a significant focus of the study.

Next Steps Sweep 9 took place when the cohort members were approximately 32 years old. Fieldwork occurred between April 2022 and September 2023. The survey used an online-first mixed-mode approach with web non-respondents being issued to CAPI interviewers after 3 weeks. Interviewers were, in addition to home visits, able to offer completion by secondary device[10], completion via LVI and in exceptional circumstances a CATI interview. The web survey also remained open throughout the F2F fieldwork period.

Ultimately, of 7,284 cohort members participating in the 9th Sweep (6,947 full completes, 337 partial completes), 85% of the responding sample participated via the online mode, secondary device was used by 2% of respondents, 2% participated via CATI, 10% were administered in-home by face-to-face interviewers, and less than 1% used MS Teams/LVI (both in-home CAPI and LVI are categorized as F2F in this study). The final completion rate, which included both full and partial completes and all eligible cohort members (n=13,820) in the calculation, was 52.7% (AAPOR Response Rate 1; The American Association for Public Opinion Research 2023). We analysed the data for the subsample of respondents who were employed at the time of the interview and provided answers to occupation questions (n=5,323).

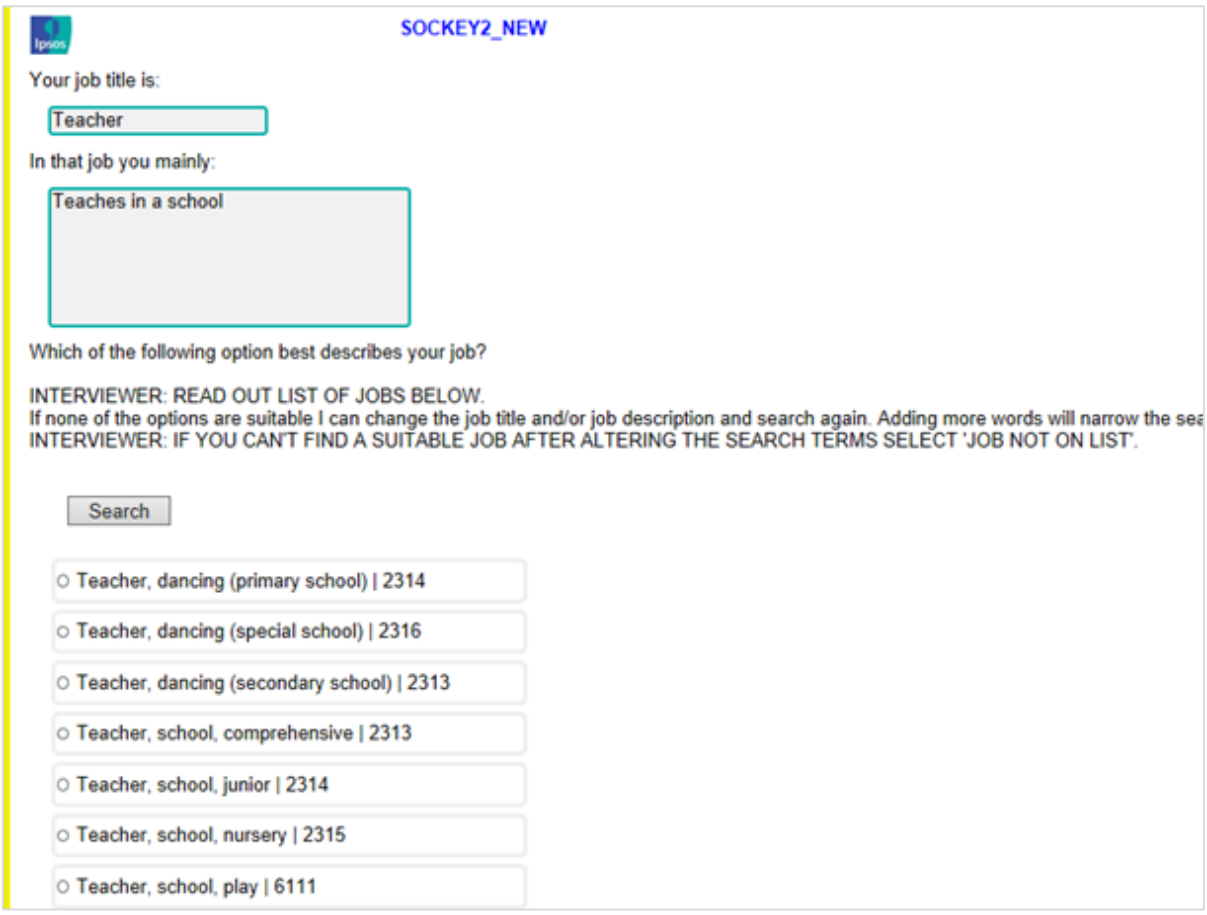
[9] In Sweeps 1-4 Next Steps collected information from parents about their occupations. Participants themselves were first asked about their occupation in Sweep 5. Across all these sweeps occupation data were collected using the "traditional" approach of collecting detailed open-ended job descriptions with subsequent office coding.

[10] A secondary device refers to a small tablet provided to participants, which interviewers later collected. Participation via a secondary device resembles online completion.

3.2 Collection of occupation data

In Sweep 8 (2015-16), a first feasibility study of the look-up approach was conducted, and the results were reported in Peycheva et al. (2021). This study used the look-up approach for all respondents and open-descriptions with subsequent office coding only if no occupation code was assigned using the look-up. In Sweep 9, a second methodological study was conducted to further investigate the quality of occupation data collected using the look-up approach. To compare the look-up coding with the traditional method of office coding open-text descriptions, all respondents were asked to use the look-up and, additionally, to provide open-descriptions of their jobs which were subsequently coded by office coders.

Figure 1: Look-up approach to collecting occupation data



Collection of occupation data in Sweep 9 using the look-up approach involved the following steps (see Figure 1 for more detail):

1) Respondents were asked *'What is your job title?'* (maximum 50 characters allowed) and *'Please tell us keywords which describe what you do in your job.'* (maximum 200 characters allowed) for their current job[11].
2) The look-up trigram search function used the provided verbatims to generate a list of possible occupations from the job title index. The list was presented to the respondent in self-administered modes or read out-loud by the interviewer in interviewer-administered modes. In-home interviewers were encouraged to hand over the tablet to respondents to help them choose the most suitable code. A maximum of 35 occupations were displayed[12].
3) The job title and job description keywords could be edited to generate a new list of occupations in the event that a suitable occupation was not generated based on the initially provided information.
4) The respondent (or interviewer after discussion with the respondent), selected the code that best described the job. In case the occupation could not be coded or selected from the list of suggested occupations, '*job not in the list'* could be chosen as the final answer to the closed-ended question.

The look-up was followed by the collection of an open-text description of the job which was subsequently office coded. Respondents were asked to describe in their own words what they mainly do in their job, with the following full question text: *"This approach to collecting information about your job is new and we are testing it out. To help us check whether it is working, could you also describe in your own words what you mainly do in your job? Please describe in detail (for example the type of work, the department you are in, and what level you work at)."*. The question was asked of all respondents, including those who successfully selected an occupation code using the look-up. This approach[13] was designed to assess the quality of the look-up approach by allowing us to compare its consistency with occupation codes collected using the "gold standard" approach.

To further explore the feasibility of occupation coding using the look-up, the following closed-ended question was asked after respondents or interviewers selected the occupation code: "*How well do you think the option you selected actually describes the job that you do?*" (answer options: 'Very well', 'Fairly well', 'Not very well', 'Not at all well'). Including this

[11] Other occupation-related questions were asked in the survey, such as a question about the partner's occupation, but only 'current occupation' data were collected using the two approaches for this methodological study.

[12] The order in which the occupation codes were displayed was based on the closeness of the match between the search string inputs and the job title index – specifically, the suggested job title with least number of whole words not included in the search string was displayed first.

[13] While the approach of asking for job title, job description keywords, and a detailed job description could potentially affect office coding (compared to a traditional approach that only asks for a detailed job description), we argue that any impact was minimal and did not affect the key findings of our study.

question enables investigation into the relationship between consistency in the look-up and office coding, and the respondent's perception of the suitability of the selected code.

3.3 Office coding of occupation data

The respondent-provided information about their jobs, presented in Table 1, was manually coded after data collection by office coders from Ipsos. They assigned a unique 4-digit code for each respondent's current occupation using the UK's Standard Occupational Classification (SOC2020) system, which is the same code-frame incorporated in the look-up function. The coders used CASCOT software to assist with the coding. This software is programmed to analyse the textual inputs and to use these to propose a list of the most likely SOC codes along with confidence scores for each of them. The software was used in a non-automated way, meaning that the coders manually reviewed the suggested SOC codes for all cases and chose what they believed to be the most suitable code. While the CASCOT's confidence score was displayed next to each of the proposed codes, the coding decision was made based on the personal judgement of the coders. To assess consistency between look-up coding and manual coding on one side, and between office coders, post-interview manual coding was conducted twice for each cohort member by two independent coders with the same level of experience. In this study, occupation codes assigned by the first office coder are compared to the look-up coding to calculate the agreement rate.

The information provided to coders slightly differed based on whether a respondent/interviewer successfully selected one of the occupation codes proposed by the look-up function, as presented in Table 1. All the information provided was fed into CASCOT, which generated a list of proposed occupations for office coders to review.

Table 1: Information provided to office coders based on whether an occupation code using the look-up approach was selected

Occupation coding data source	Information provided to office coders	Scenario	
		Look-up code selected	Look-up code not selected
Look-up	Job title *(from the initial search)*	✓	✓
	Final job title *(from the final search)*	✗	✓
	Keyword job description *(from the initial search)*	✗	✓
	Final keyword job description *(from the final search)*	✗	✓
Open-description	Detail job description *(a separate open-ended request)*	✓	✓
Additional items (separate questions)	Special qualifications required to do the job (if any)	✓	✓
	Main product of firm or organization	✓	✓

3.4 Statistical analysis

We employ Chi-square testing to examine differences in coding rates and agreements between the coding approaches, as well as between the modes of survey data collection. Additionally, binary logistic regression analysis is used to analyse various factors affecting the following as our key binary outcome variables:

- occupation coding outcomes (occupation code selected vs. occupation code not selected),
- the agreement between the look-up and office coding, and
- the agreement between two coders (same occupation code selected vs. different occupation code selected).

We use a range of predictors which could be categorized into four distinct groups: (i) survey administration (mode and device), (ii) look-up input metrics (e.g. length of keywords, time spent using the look-up - measured with paradata), (iii) socio-demographics, and (iv) respondent reported suitability of the selected code.

For categorical predictors, we generally select the reference category with the largest relative frequency among all categories for that variable. To address the issue of non-linear relationships, we recode the lengths of initial recorded answers to the three job information questions (i.e., job title, job keywords, and the open-ended description of jobs) into five groups: 0 characters will constitute the first category, and the remainder of the sample will be recoded into four groups of approximately equal sizes. Timing variables are top-coded, with all values larger than the 99th percentile replaced by the 99th percentile value.

To test for the significance of multiple marginal effects, we use the Wald test, which is recommended for nonlinear regressions such as binary logistic regression analysis (Mize 2019). SPSS 29.0 was used for data processing and Stata/SE 17.0 for data manipulation and all statistical analyses, including regression modelling.

4. Results

4.1 Occupation coding rates

Look-up and office coding rates. Rates for the selection of occupation codes are, as explained in the Introduction, our first indicator of the feasibility of the look-up approach to occupation coding. We address the first research question on occupation coding rates (RQ1) by examining any differences between the look-up approach and the open-description with office coding approach, and between the three main groups of survey modes: Web (including secondary device), F2F (including in-home and LVI), and CATI. We also examine how accurately respondents felt the selected look-up codes described their occupation (see Table 2) and explore factors affecting the look-up coding rates (see Table 3).

Open-ended job descriptions could be office-coded for the vast majority of respondents, with no statistically significant differences between the modes (Web: 99.3%, F2F: 99.5%, CATI: 100%, see Table 2). SOC codes were assigned by office coders for all but 32 respondents (out of 5,323), including the majority of those who did not select an occupation code using the look-up. Compared to office coding, the proportion of respondents who successfully selected an occupation code using the look-up was notably lower. The coding rate was 81.5% in the Web mode. The rest of the sample (18.5%) either selected 'job not in the list', answered 'don't know', refused to provide any textual information, or declined to choose an occupation code from the list. The proportions of respondents with a valid look-up occupation code were higher in both interviewer-administered modes (F2F: 88.2%, CATI: 93.0%).

Table 2: Coding rates for office coding and look-up coding, and self-reported accuracy of look-up coding, by survey mode

Mode	Office coding	Look-up			
	Coding rate	*Coding rate*	*Self-reported accuracy of occupation coding**		
			Very well %	Fairly well %	Not very/at all well** %
Web, including secondary device (n=4,626)	99.4%	81.5%[bc]	44.8%[bc]	49.3%	5.9%
F2F, including in-home and LVI (n=583)	99.5%	88.2%[a]	62.3%[a]	36.0%	1.7%
CATI (n=114)	100.0%	93.0%[a]	58.5%[a]	36.8%	4.7%
***Total** (n=5,323)*	***99.4%***	***82.5%***	***47.4%***	***47.3%***	***5.3%***

*[a b c] indication of statistically significant differences between the groups (a=Web, b=F2F, c=CATI) at p<0.05 (Chi-Square test); [*]only those who selected an occupation code using the look-up, were asked the follow-up question; **due to a very small proportion of respondents selecting 'Not at all well', we combined the groups 'Not very well' and 'Not at all well'*

When asked to rate the accuracy of the selected code nearly 95% of all respondents selected answers 'Very well' or 'Fairly well'. There were statistically significant differences between the modes with the proportion of respondents who felt the code described their occupation 'Very well' being higher in both F2F (62.3%) and CATI (58.5%) modes compared to the Web mode (44.8%). These results confirm that not only did a smaller proportion of Web respondents select an occupation code using the look-up, but those who did were, on average, less confident in the adequacy of the code.

Factors affecting the look-up coding rates. To extend the analysis conducted to study coding rates and address RQ1, we carry out a logistic regression analysis to examine which look-up input metrics, mode and device, as well as personal characteristics (see Tables A1 and A2 in Appendix A for descriptive statistics), explain the occupation coding rates. Therefore, in addition to analysing the impact of mode and device on coding rates, we gradually include a range of look-up procedural (see Models 1-3 in Table 3) and socio-demographic predictors (see Table B1 in Appendix B).

The results from Model 1 confirm our previous findings, indicating that occupation coding rates were higher in interviewer-administered modes, namely F2F (including LVI) and CATI. While there were no observable differences in coding rates between PC/laptop, tablet, and smartphone devices, respondents for which no information about the self-completion device was recorded in the data had slightly lower coding rates. However, the effect of mode was mitigated with the addition of the look-up input metrics predictors to the logistic regression modelling, and there is no longer any statistically significant impact of mode/device in Model 3 with the largest Pseudo R-Square values of all models. In other words, input metrics associated with survey administration are able to explain why using the F2F and CATI modes resulted in higher occupation coding rates.

Moreover, the coefficients for Model 2 show that coding rates are positively affected by the length of the entered job title and negatively affected by the length of entered keywords, the time spent entering the job title, and editing the job information to regenerate the list of offered occupation codes (consistent with the Model 3 results). While the length of job information variables was included as numeric predictors in Model 2, we later observed that the association was not, in fact, linear. Therefore, Model 3 includes these variables in a categorical form, with the results supporting this change. The coefficients explain that not entering job information[14] negatively impacted coding rates. An occupation code was most likely to be selected by respondents with job title entries between 14 and 25 characters and keywords describing the job between 1 and 37 characters in length. Importantly, longer keyword descriptions (i.e., 38+ characters) had a negative impact on coding rates. Lastly, the coefficients for the office-coded SOC code from this model show that occupation coding rates differed between 1-digit SOC groups. Compared to 'Professional Occupations' as the reference group which was the largest in size, 'Managers, directors and senior officials' and 'Administrative and secretarial occupations' had lower look-up coding rates[15].

[14] This was recorded as zero characters for the length of the job title or keywords. The lookup produced a list of suggested occupation codes even if either the job title or keywords were not provided, but it did not produce suggestions if no job information was provided.

[15] Additionally, we investigated the effect of socio-demographics on occupation coding (see Table B1, outcome variable: 'Occupation code selected in the look-up'). The results confirm the findings based on Model 3, but they show no impact of sex, highest qualification, relative deprivation, household size, and housing tenure on coding rates. On the other hand, respondents residing in the UK states other than England and those who are married or in a civil partnership were more likely to select a valid occupation code using the look-up, while people of South Asian ethnicity (i.e., Indian, Pakistani, Bangladeshi) were less likely to select a valid occupation code.

Table 3: Logistic regression analysis results, binary outcome variable: *occupation code selected in the look-up*

Predictors	Model 1 (n=5,305)	Model 2 (n=5,296)	Model 3 (n=5,267)
	Coef.	Coef.	Coef.
Mode & device: Web, tablet	-0.00	0.00	0.11
Mode & device: Web, smartphone	-0.04	-0.11	-0.11
Mode & device: Web, not defined	-0.33*	-0.32*	-0.29
Mode & device: F2F (including LVI)	0.47**	0.39*	0.28
Mode & device: CATI	1.04**	0.95*	0.72
Length of job title: numeric		0.03***	
Length of job title: 0 characters			-3.89***
Length of job title: 14-18 characters			0.41***
Length of job title: 19-25 characters			0.51***
Length of job title: 26-50 characters			0.29*
Length of keywords describing job: numeric		-0.002*	
Length of keywords describing job: 0 char.			-1.35***
Length of keywords describing job: 20-37 char.			-0.22
Length of keywords describing job: 38-65 char.			-0.39***
Length of keywords describing job: 66-200 char.			-0.54***
Timing, job title: numeric		-0.01***	-0.01**
Timing, job keywords: numeric		-0.00	-0.00
Edited job information entries: Yes		-0.68***	-0.40***
Same job as in previous sweep: Yes[a]		0.10	0.15
SOC: 1 Managers, directors and senior officials			-0.48***
SOC: 3 Associate professional occupations			-0.21
SOC: 4 Administrative and secretarial occupations			-0.44**
SOC: 5 Skilled trades occupations			-0.21
SOC: 6 Caring, leisure and other service occupations			-0.19
SOC: 7 Sales and customer service occupations			-0.19
SOC: 8 Process, plant and machine operatives			-0.27
SOC: 9 Elementary occupations			-0.32
Constant	*1.54****	*1.46****	*2.14****
Pseudo R-Square (McFadden R2)	*0.007*	*0.031*	*0.081*

[a] this binary variable needed to be included as a control variable because respondents with the same job as in the previous sweep were not asked to enter their job title, resulting in a length of job title equal to 0; Reference categories: Mode & device: Web, PC/laptop, Length of job title: 1-13 characters, Length of keywords describing job: 1-19 characters, SOC: 2 Professional occupations; ***p<0.001, **p<0.01, *p<0.05

4.2 Agreement in coding

Look-up – office coding and inter-coder agreement rates. Our second indicator of the quality of the look-up approach to occupation coding was agreement between the codes selected from the list generated by the look-up function and those assigned by office coders. Agreement rates could only be calculated for those having both codes selected in the look-up and assigned by office-coders (n= 4,388), effectively having to exclude 17.6% of the sample due to an absence of either of the codes. We address the second research question on occupation coding agreement (RQ2) by examining the differences between the look-up – office coding agreement and the agreement between two office coders. We also focus on identifying any differences between the three main groups of survey modes and examine factors affecting coding agreement. The results are presented in Tables 4 and 5.

Table 4: Agreement between look-up and office coding, and between two office coders, at SOC 1-digit to 4-digit levels, by survey mode

Mode	Agreement the look-up & office coder[16]				Agreement 1st office coder & 2nd office coder			
	1-digit level	2-digit level	3-digit level	4-digit level	1-digit level	2-digit level	3-digit level	4-digit level
Web, including secondary device (n=3,768)	78.3%	74.1%	69.9%	62.1%	94.3%	93.1%	92.2%	89.2%
F2F, including in-home and LVI (n=514)	77.6%	73.5%	68.7%	62.8%	95.2%	94.5%	93.5%	91.6%
CATI (n=106)	81.1%	79.3%	75.5%	69.8%	93.0%	92.1%	91.2%	88.6%
Total (n=4,388)	***78.3%***	***74.2%***	***69.9%***	***62.3%***	***94.4%***	***93.3%***	***92.3%***	***89.4%***

The evidence reveals a notable and statistically significant difference in agreement rates between (i) the look-up SOC code and the SOC code assigned by an office coder, and (ii) between two office coders, across all SOC levels. The greatest difference in agreement rates for the entire sample is observed at the 4-digit level and the smallest at the 1-digit level. In contrast, no statistically significant differences are observed between the survey modes, whether in the agreements between the look-up code and the code assigned by an office coder, or between the two office coders, with these findings consistent across all four SOC levels. The agreement rates between the look-up and office coding decrease gradually as additional digits are added to the code. The decline in agreement rates between two coders at different SOC levels is similarly gradual but much less pronounced; for example, there is

[16] We are presenting the agreement rates between the look-up code and the office coder who coded the job information for a respondent first. It is worth noting that, due to the very high agreement between the office coders, the agreement rates between the look-up code and the second office coder are very similar to those between the look-up code and the first office coder, and the findings are the same.

only about a 3-percentage-point decrease in agreement rates from the 1-digit level to the 4-digit level for F2F respondents.

Factors affecting the 'look-up – office coding' agreement rates. To further address RQ2, we carry out a regression analysis with 'SOC 4-digit look-up – office coding agreement' as the binary outcome variable. Look-up input metrics, mode and device, self-reported suitability of the selected code, and personal characteristics are gradually added as predictors to Models 1-3 (see Table 5).

The findings on the effect of mode and device on agreement rates are similar to those on the effect of mode and device on the look-up occupation coding rates. The results from Model 1 show differences between Web PC/Laptop, Web smartphone, and CATI, but the addition of the look-up input metrics nullifies this effect in Models 2-3. Interestingly, the length of job title, editing job information, timings, or length of open descriptions of jobs (as a question for office coding only) had no effect on agreement rates in Model 2. However, longer keywords describing jobs (20+ characters) had a negative impact on agreement rates, similar to the effect on the look-up coding rates.

This is confirmed by the results from Model 3, where we observe an additional effect of office-coded occupation groups, which is fairly consistent with our findings for occupation coding rates. Also, the agreement rates decline significantly as the selected answer moves further from the first choice, indicating that the further away an answer is from the top selection, the lower the agreement. Lastly, the respondent-reported accuracy of the selected occupation code proved to be a very strong predictor of the look-up – office coding agreement, with respondents selecting 'Very well' having much higher agreement rates compared to those selecting 'Fairly well' and especially those selecting either 'Not very well' or 'Not at all well'[17,18].

[17] To provide findings on the effect of socio-demographics on agreement rates, we added them to the predictors from Model 3 and presented them in a separate model in Table B1 (outcome variable: 'Look-up – office coding agreement'). In addition to confirming the results from Model 3, the analysis identified highest qualification as the only socio-demographic predictor having an effect on the outcome variable, with the groups of respondents with a postgraduate degree and A-level or equivalent having higher agreement rates than those with a graduate degree.

[18] As supplementary analysis, we looked at the effects of the same range of predictors as for the look-up-office coding agreement on the agreement between two coders, which are also presented in Table B1. As expected, the inter-coder agreement is negatively affected by respondents providing no open descriptions. At the same time, lower inter-coder agreement is also negatively associated with respondents selecting one of the last offered occupation codes in the look-up (i.e., 13th-35th) and longer job titles (i.e., more than 25 characters).

Table 5: Logistic regression analysis results, binary outcome variable: *SOC 4-digit coding look-up – office coding agreement*

Predictors	Model 1 (n=4,379)	Model 2 (n=4,372)	Model 3 (n=4,090)
	Coef.	Coef.	Coef.
Mode & device: Web, tablet	0.18	0.10	0.10
Mode & device: Web, smartphone	0.16*	0.04	0.03
Mode & device: Web, not defined	0.24	0.08	0.08
Mode & device: F2F (including LVI)	0.17	-0.02	-0.24
Mode & device: CATI	0.47*	0.19	0.20
Length of job title: 14-18 characters		0.00	-0.03
Length of job title: 19-25 characters		-0.07	-0.05
Length of job title: 26-50 characters		-0.16	-0.03
Length of keywords describing job: 0 char.		0.25	0.40
Length of keywords describing job: 20-37 char.		-0.27**	-0.22*
Length of keywords describing job: 38-65 char.		-0.48***	-0.32**
Length of keywords describing job: 66-200 char.		-0.67***	-0.56***
Length of open-descriptions of jobs: 0 characters		-0.04	-0.12
Length of open-descriptions of jobs: 39-75 characters		-0.11	-0.07
Length of open-descriptions of jobs: 76-143 characters		-0.14	-0.04
Length of open-descriptions of jobs: 144-200 characters		-0.04	0.01
Timing, job title: numeric		-0.00	-0.00
Timing, job keywords: numeric		-0.00	0.00
Edited job information entries: Yes		-0.17	0.02
SOC: 1 Managers, directors and senior officials			-0.56***
SOC: 3 Associate professional occupations			-0.37***
SOC: 4 Administrative and secretarial occupations			-0.40**
SOC: 5 Skilled trades occupations			-0.08
SOC: 6 Caring, leisure and other service occupations			0.06
SOC: 7 Sales and customer service occupations			-0.24
SOC: 8 Process, plant and machine operatives			-0.09
SOC: 9 Elementary occupations			0.30
Look-up answer selected: 2nd answer[a]			-0.31**
Look-up answer selected: 3rd-5th answer			-0.60***
Look-up answer selected: 6th-12th answer			-0.87***
Look-up answer selected: 13th-35th answer			-1.16***
Suitability of look-up code: Very well			0.73***
Suitability of look-up code: Not very/at all well			-0.84***
Constant	*0.37****	*0.99****	*1.19****
Pseudo R-Square (McFadden R^2)	*0.001*	*0.016*	*0.096*

[a]answers were not ordered alphabetically, but rather by the likelihood of the occupation code, determined based on the job description; Reference categories: Mode & device: Web, PC/Laptop, Length of job title: 1-13 characters, Length of keywords describing job: 1-19 characters, Length of open-descriptions of jobs: 1-38 characters, SOC: 2 Professional occupations, Look-up answer selected: 1st answer, Suitability of look-up code: Fairly well; ***p<0.001, **p<0.01, *p<0.05

5. Discussion and conclusion

This study provides valuable evidence in the field of occupation coding, particularly regarding coding during interviews using the look-up method. It uses survey data collected in a longitudinal study in the UK, which used both look-up coding and manual office coding for the entire sample to collect occupation data. This double-coding approach facilitates a comparison between the quality and feasibility of these two coding methods and identifies differences between self-completion and interviewer administration. It expands the existing literature (e.g., Peycheva et al. 2021; Schierholz et al. 2018; Simson et al. 2023; Tijdens 2015) by examining two key indicators of office coding feasibility: coding rates and agreement rates, specifically between look-up and office coding and between two office coders. The findings offer guidance on applying look-up functions in other studies, identifying potential solutions for improving coding and agreement rates, and demonstrating how and when to combine look-up coding with office coding for the best possible outcomes.

The look-up coding rates in our study were higher for the F2F and CATI modes compared to past studies (Hacking et al. 2006; Peycheva et al. 2021; Schierholz et al. 2018), but lower for the Web mode when compared to the study collecting data from the same Next Steps cohort at age 25 (Peycheva et al. 2021). Additionally, one of the central findings of this study is that the look-up occupation coding rates, as our first indicator of the feasibility of the look-up approach, differ between interviewer-administered (F2F, CATI) and self-administered modes (Web). The evidence suggests that self-completion leads to lower coding rates (about 82%) than interviewer administration (around 90%), which is contrary to some previous evidence (Peycheva et al. 2021)[19]. The presence of an interviewer not only led to higher coding rates but also to higher perceived accuracy of the selected code. It also appears to affect various look-up input metrics, which ultimately explain the differences in coding rates between the modes, some of which were previously reported in the literature (cf. Conrad et al. 2016; Hacking et al. 2006; Peycheva et al. 2021; Schierholz et al. 2018). These aspects include not entering job titles or keywords describing jobs, or entering rather lengthy keywords, which might not have happened in the presence of an interviewer. After controlling for those aspects, the effect of the mode is no longer statistically significant[20].

The other main finding of this study is that the agreement between the look-up codes and those assigned by an office coder, which was as low as 62% at the 4-digit SOC level, is much lower than the agreement between two office coders, especially at the 3- and 4-digit levels (about 90%). This finding is inconsistent with the results reported by Schierholz et al. (2018) for occupation coding in a telephone survey. Moreover, we did not observe any differences between survey modes in agreement rates after covariate adjustment, even though respondents participating via interviewer-administered modes reported higher perceived

[19] That said, we must acknowledge certain differences in complex sample design, survey weights, and the look-up algorithm between the compared studies.

[20] Some of these procedural aspects could be managed in the look-up search function in the self-completion mode, such as with prompts or additional instructions, at least to a certain extent.

accuracy of the selected code. If manual coding with the use of occupation coding software is still considered the "gold standard", the difference in agreement rate between look-up codes and office coding compared with between two office coders could be interpreted as evidence that the occupation data collected via look-up is of poorer quality. However, this discrepancy may be better explained by very high inter-coder agreement rates rather than lower look-up–office coding agreement rates. The high inter-coder agreement can likely be strongly attributed to the two office coders using exactly the same methodology. They had received the same standardized training and used the same software, which used the same algorithm to generate suggested codes using the same text inputs. In contrast, the look-up and office coding were conducted by different people (i.e., respondents, interviewers, coders) using different text inputs. In a different set-up, where no coding software was used, or if two independent coders from separate organizations applied different approaches (as in the study by Schierholz et al. 2018), inter-coder agreement would likely be much lower and could potentially be comparable to the look-up–office coding agreement, as observed in similar studies.

Additionally, we identified several different factors that negatively affect the agreement between look-up and office coded codes. The most notable ones were the self-perceived accuracy of the code and the ranking of the suggested codes, which were key aspects not explored previously in similar research on occupation coding. The length of the keywords describing jobs also had a negative effect on both coding and agreement rates, which is consistent with findings from several previous studies (Conrad et al. 2016; Helppie-McFall and Sonnega 2018; Massing et al. 2019). These aspects measured with the look-up input metrics could, with decent accuracy, predict discrepancies between the codes in advance and allow for potential adjustments in the occupation coding data collection and coding procedures.

Considering both promising look-up occupation rate results and slightly less promising agreement results, we suggest that coding during the interview should be supplemented with subsequent office coding (see Burstyn et al. 2014), and this recommendation does not differ between the survey modes. The findings from our study, as well as previous research (e.g., Schierholz et al. 2018), have shown that in cases where an occupation code is not selected by the respondent during the interview, office coders can successfully code occupations for a vast majority of those respondents with a missing occupation code. Our evidence suggests that, with respect to cost implications, manual coding would ideally be conducted for a subsample and not for the whole sample, as in our study. We would recommend using additional office coding in certain cases for a data quality check. Those include instances where respondents have to edit their entries and regenerate a list of occupation codes, when they do not select one of the first answers from the proposed list, when they report low confidence in the selected code, or for particular groups of occupations (roughly estimated to about 15% of the sample).

Additionally, this evidence suggests that certain technical solutions might make occupation coding during the interview more feasible. It appears that for certain occupations, the look-up function was less likely to generate an accurate list of possible occupations to the respondent. Hence, another solution would be to explore revising the response options in the

look-up. Also, the look-up function might not work as well with longer descriptions of jobs, which resulted in lower coding rates and coding agreements; albeit this could also be explained by occupations that are more challenging to self-code, requiring complex job descriptions that provide more occupational detail. Since the evidence from this study cannot fully explain this phenomenon which has been reported previously, further research on the technical aspects of the look-up is required. This research would ideally also include testing different prompts to respondents in the self-completion mode in cases of the look-up input events that lead to lower coding rates, which might ultimately close or narrow the gap between self-completion and interviewer administration.

Lastly, we must recognize certain limitations of this research. As respondents were not randomly allocated to survey modes, we cannot exclude the effect of compositional differences that might be associated with the outcome variables, even after controlling for a range of individual, including occupational characteristics. A fully randomized survey experiment or the use of other methods for causal inference could make the findings more robust. Additionally, when using agreement rates as an indicator of the feasibility of the look-up, we assume that office coding remains the "gold standard" in occupation coding. This assumption might need to be challenged, as there could be cases where respondents have a more comprehensive knowledge and understanding of their own occupation than office coders, who are asked to make coding decisions based on short or inaccurate descriptions of other people's jobs. Further research using a combination of coding approaches, including the newest automatic coding methods and language-based models, might offer additional evidence on this issue. Lastly, although the look-up data reveal certain procedural details measured with the look-up input metrics at an individual level, there might be other aspects that could better explain how respondents interact with the look-up and how the function could be improved for higher coding efficiency and accuracy. Cognitive interviewing combined with usability testing might provide valuable insights into this issue. Nonetheless, the evidence from our study shows promising results for integrating occupation coding during the interview and indicates in what scenarios look-up coding should be combined with office coding to achieve a balance between cost-efficiency and occupation data quality.

References

The American Association for Public Opinion Research. (2023), *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*, 10th edition. AAPOR.

Belloni, M., Brugiavini, A., Meschi, E., and Tijdens, K. (2016), "Measuring and Detecting Errors in Occupational Coding: An Analysis of SHARE Data," *Journal of Official Statistics*, 32(4), 917-945.

Brugiavini, A., Belloni, M., Martens, M., and Buia, R. E. (2017), "The 'Job Coder'," in SHARE Wave 6: Panel Innovations and Collecting Dried Blood Spots (pp. 51-70). MEA, Max Planck Institute for Social Law and Social Policy.

Burstyn, I., Slutsky, A., Lee, D. G., Singer, A. B., An, Y., and Michael, Y. L. (2014), "Beyond Crosswalks: Reliability of Exposure Assessment Following Automated Coding of Free-Text Job Descriptions for Occupational Epidemiology," *Annals of Occupational Hygiene*, 58(4), 482-492.

Conrad, F. G., Couper, M. P., and Sakshaug, J. W. (2016), "Classifying Open-Ended Reports: Factors Affecting the Reliability of Occupation Codes," *Journal of Official Statistics*, 32(1), 75-92.

Dillman, D. A., Smyth, J. D., and Christian, L. M. (2014), *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. John Wiley and Sons.

Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M., and Steiner, S. (2017), "Three Methods for Occupation Coding Based on Statistical Learning," *Journal of Official Statistics*, 33(1), 101-122.

Hacking, W., Michiels, J., and Janssen-Jansen, S. (2006), "Computer Assisted Coding by Interviewers," In Proc. 10th Int. Blaise Users Conf (pp. 283-296).

Helppie-McFall, B., and Sonnega, A. (2018), "Feasibility and Reliability of Automated Coding of Occupation in the Health and Retirement Study," Michigan Retirement Research Center Research Paper, (2018-392).

Herzing, J. M. (2020), "Investigation of Alternative Interface Designs for Long-List Questions–The Case of a Computer-Assisted Survey in Germany," *International Journal of Social Research Methodology*, 23(6), 639-650.

Hoffmann, E., and Thomas, R. (1995), "What Kind of Work Do You Do? Data Collection and Processing Strategies When Measuring 'Occupation' for Statistical Surveys and Administrative Records," International Labour Organization.

International Labour Organization. (2010), https://www.ilo.org/public/english/bureau/stat/isco/

Kim, C., Kim, J., and Ban, M. (2020), "Do You Know What You Do for a Living? Occupational Coding Mismatches Between Coders in the Korean General Social Survey," *Research in Social Stratification and Mobility*, 70, 100467.

Lyberg, L., and Dean, P. (1992), "Automated Coding of Survey Responses: An International Review, " R&D Reports (1992–2). Statistics Sweden, Stockholm, Sweden.

Mannetje, A. T., and Kromhout, H. (2003), "The Use of Occupation and Industry Classifications in General Population Studies," *International Journal of Epidemiology*, 32(3), 419-428.

Massing, N., Wasmer, M., Wolf, C., and Zuell, C. (2019), "How Standardized is Occupational Coding? A Comparison of Results from Different Coding Agencies in Germany," *Journal of Official Statistics*, 35(1), 167-187.

Mize, Trenton D. (2009), "Best practices for estimating, interpreting, and presenting nonlinear interaction effects, " *Sociological Science*, 6, 81-117.

Office for National Statistics. (n.d.). “Standard Occupational Classification (SOC), ” [online] Available at https://www.ons.gov.uk/methodology/classificationsandstandards/standardoccupationalclassificationsoc.

Ossiander, E. M., and Milham, S. (2006), “A Computer System for Coding Occupation,” *American Journal of Industrial Medicine*, 49(10), 854-857.

Peycheva, D. N., Sakshaug, J. W., and Calderwood, L. (2021), “Occupation Coding During the Interview in a Web-First Sequential Mixed-Mode Survey,” *Journal of Official Statistics*, 37(4), 981-1007.

Russ, D. E., Josse, P., Remen, T., Hofmann, J. N., Purdue, M. P., Siemiatycki, J., Silverman, D. T., Zhang, Y., Lavoué, J., and Friesen, M. C. (2023), “Evaluation of the Updated SOCcer v2 Algorithm for Coding Free-Text Job Descriptions in Three Epidemiologic Studies,” *Annals of Work Exposures and Health*, 67(6), 772-783.

Safikhani, P., Avetisyan, H., Föste-Eggers, D., and Broneske, D. (2023), “Automated Occupation Coding with Hierarchical Features: A Data-Centric Approach to Classification with Pre-Trained Language Models,” *Discover Artificial Intelligence*, 3(1), 6.

Schierholz, M., Gensicke, M., Tschersich, N., and Kreuter, F. (2018), “Occupation Coding During the Interview,” *Journal of the Royal Statistical Society Series A: Statistics in Society*, 181(2), 379-407.

Schierholz, M., and Schonlau, M. (2021), “Machine Learning for Occupation Coding—A Comparison Study,” *Journal of Survey Statistics and Methodology*, 9(5), 1013-1034.

Simson, J., Kononykhina, O., and Schierholz, M. (2023), “occupationMeasurement: A Comprehensive Toolbox for Interactive Occupation Coding in Surveys,” *Journal of Open Source Software*, 8(88), 5505.

Tijdens, K. G. (2014), “Reviewing the Measurement and Comparison of Occupations Across Europe,” AIAS Working Paper.

Tijdens, K. (2015), “Self-Identification of Occupation in Web Surveys: Requirements for Search Trees and Look-Up Tables,” *Survey Methods: Insights from the Field*, 1-11. https://doi.org/10.13094/SMIF-2015-00008.

Tijdens, K. (2016), “Measuring Occupations: Respondent’s Self-Identification from a Large Database.”

Tijdens, K. (2022), “The Importance of Occupation Coding Quality: Lessons for EU-SILC from SHARE and Other International Surveys,” in *Improving the Measurement of Poverty and Social Exclusion in Europe: Reducing Non-Sampling Errors*, eds. P. Lynn, and L. Lyberg, Luxembourg: Publications Office of the European Union, pp. 413-426.

U.S. Bureau of Labor Statistics. (n.d.), *Standard Occupational Classification,* [online]. Available at https://www.bls.gov/soc/.

Appendix A

Table A1: Descriptive statistics – categorical independent variables (n=5,323)

Variable	Variable category	Frequency	Relative frequency
Mode & device	Web, PC/Laptop	1,294	24.4%
	Web, tablet	187	3.5%
	Web, smartphone	2,774	52.3%
	Web, not defined	353	6.7%
	F2F (including LVI)	583	11.0%
	CATI	114	2.1%
Length of job title	0 characters	97	1.8%
	1-13 characters	1,283	24.1%
	14-18 characters	1,418	26.6%
	19-25 characters	1,336	25.1%
	26-50 characters	1,189	22.3%
Length of keywords describing job	0 characters	166	3.1%
	1-19 characters	1,287	24.2%
	20-37 characters	1,302	24.5%
	38-65 characters	1,295	24.3%
	66-200 characters	1,273	23.9%
Length of open-descriptions of jobs	0 characters	612	11.5%
	1-38 characters	1,177	22.1%
	39-75 characters	1,157	21.7%
	76-143 characters	1,203	22.6%
	144-200 characters	1,174	22.1%
Edited job information entries	No	4,561	85.7%
	Yes	762	14.3%
Same job as in previous sweep	No	4,589	86.2%
	Yes	734	13.8%
Standard Occupational Classification (office coded)	1 Managers, directors and senior officials	669	12.6%
	2 Professional occupations	1,675	31.7%
	3 Associate professional occupations	1,022	19.3%
	4 Administrative and secretarial occupations	507	9.6%
	5 Skilled trades occupations	356	6.7%
	6 Caring, leisure and other service occupations	399	7.5%
	7 Sales and customer service occupations	264	5.0%
	8 Process, plant and machine operatives	176	3.3%
	9 Elementary occupations	223	4.2%
Look-up answer selected[a]	1st answer	1,687	38.5%
	2nd answer	613	14.0%
	3rd-5th answer	789	18.0%
	6th-12th answer	666	15.2%
	13th-35th answer	630	14.4%
Suitability of look-up code[a]	Very well	1,946	47.4%
	Fairly well	1,943	47.3%
	Not very well	208	5.1%
	Not at all well	11	0.3%
Sex	Male	2,191	46.1%
	Female	2,562	53.9%

Highest qualification	Level 1 - GCSE lower grades or equivalent	238	4.6%
	Level 2 - GCSE higher grades or equivalent	584	11.2%
	Level 3 - A-level or equivalent	621	11.9%
	Level 4 - Graduate degree	2,288	44.0%
	Level 5 - Postgraduate degree	1,228	23.6%
	Other academic qualifications	242	4.7%
Region	North East England	215	4.1%
	North West England	698	13.3%
	Yorkshire and the Humber	545	10.4%
	East Midlands	447	8.5%
	West Midlands	617	11.7%
	East of England	565	10.7%
	London	903	17.2%
	South East	770	14.6%
	South West	429	8.2%
	Other UK states	73	1.4%
Index of Multiple Deprivation	1st decile	501	9.5%
	2nd decile	536	10.2%
	3rd decile	596	11.3%
	4th decile	571	10.9%
	5th decile	550	10.5%
	6th decile	500	9.5%
	7th decile	515	9.8%
	8th decile	524	10.0%
	9th decile	515	9.8%
	10th decile	446	8.5%
Household size	1 person	629	11.8%
	2 persons	1,678	31.5%
	3 persons	1,419	26.7%
	4 persons	1,075	20.2%
	5 or more persons	522	9.8%
Housing tenure	Own outright	284	5.4%
	Own, buying with help of mortgage/loan	2,700	51.1%
	Rent it	1,460	27.6%
	Live rent free	393	7.4%
	Other	450	8.5%
Marital status	Married or a civil partner	2,228	42.0%
	Single (never married)	2,943	55.5%
	Divorced	111	2.1%
	Other	21	0.4%
Ethnicity	White	3,759	71.3%
	Mixed	267	5.1%
	South Asian	794	15.1%
	Black	359	6.8%
	Other	97	1.8%

[a] *can be reported only for respondents who selected an occupation code in the look-up*

Table A2: Descriptive statistics – numeric independent variables

	n	Mean	Median	Std. dev.	Min	Max[a]
Length of job title (no. of characters)	5,323	19.54	18.00	9.63	0	56
Length of keywords describing job (no. of characters)	5,323	48.16	36.00	41.66	0	200
Length of open-descriptions of jobs (no. of characters)	5,323	80.41	65.00	64.11	0	200
Timing, job title (seconds)	5,316	13.03	10.00	13.88	0	92
Timing, job keywords (seconds)	5,316	35.43	22.00	48.01	1	361
Look-up answer selected (no. of answer)[b]	4,385	5.81	2.00	7.65	1	35

[a] timing variables were top-coded (all values larger than the 99th percentile were replaced by the 99th percentile value); [b] can be reported only for respondents who selected an occupation code in the look-up; Std. dev. – standard deviation

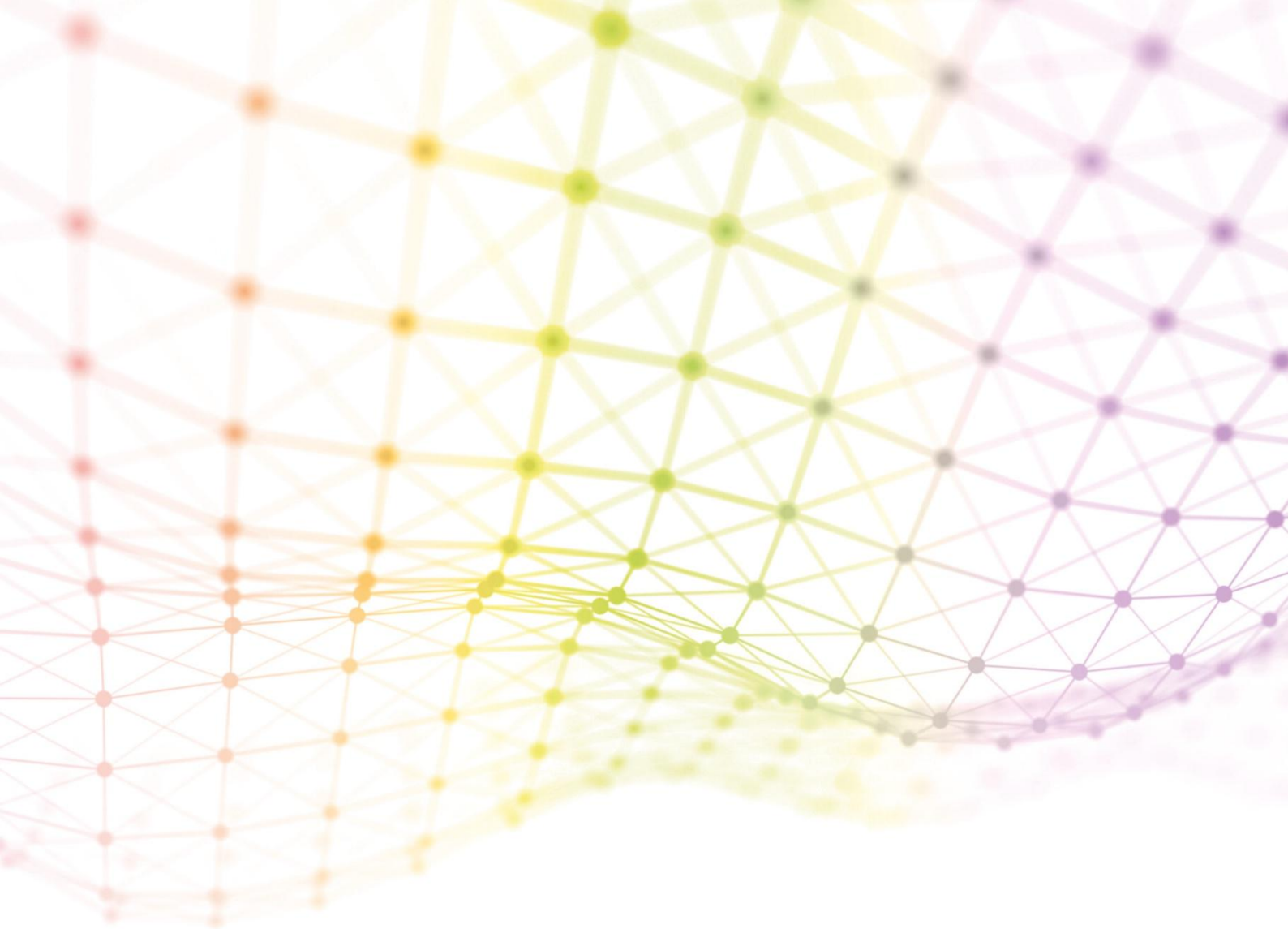
Appendix B

Table B1: Logistic regression analysis results with socio-demographic predictors, binary outcome variables: *occupation code selected in the look-up, SOC 4-digit coding look-up – office coding agreement, SOC 4-digit agreement between two coders*

Predictors	**Model 5: Occupation code selected in the look-up (n=4,613)**	**Look-up – office coding agreement (n=3,839)**	**Agreement between two coders (n=3,839)**
	Coef.	Coef.	Coef.
Mode & device: Web, tablet	-0.06	0.13	0.34
Mode & device: Web, smartphone	-0.13	0.00	0.16
Mode & device: Web, not defined	-0.09	0.05	-0.26
Mode & device: F2F (including LVI)	0.27	-0.23	0.09
Mode & device: CATI	0.56	0.58*	-0.15
Length of job title: 0 characters	-4.31***	/	/
Length of job title: 14-18 characters	0.27*	0.00	-0.12
Length of job title: 19-25 characters	0.42**	-0.05	-0.16
Length of job title: 26-50 characters	0.18	-0.06	-0.48**
Length of keywords describing job: 0 characters	-1.44***	0.63	1.56
Length of keywords describing job: 20-37 characters	-0.23	-0.21	-0.08
Length of keywords describing job: 38-65 characters	-0.41**	-0.34**	-0.01
Length of keywords describing job: 66-200 characters	-0.56***	-0.58***	-0.24
Length of open-descriptions of jobs: 0 characters		0.04	-0.91***
Length of open-descriptions of jobs: 39-75 characters		-0.08	-0.04
Length of open-descriptions of jobs: 76-143 characters		-0.05	-0.06
Length of open-descriptions of jobs: 144-200 characters		0.05	0.21
Timing, job title: numeric	-0.01	0.00	-0.00
Timing, job keywords: numeric	-0.002*	0.00	0.00
Edited job information entries: Yes	-0.46***	0.04	-0.01
Same job as in previous sweep: Yes	0.17		
SOC: 1 Managers, directors and senior officials	-0.59***	-0.52***	-0.28
SOC: 3 Associate professional occupations	-0.30	-0.36**	-0.02
SOC: 4 Administrative and secretarial occupations	-0.50**	-0.39	-0.38
SOC: 5 Skilled trades occupations	-0.43*	-0.19	0.01
SOC: 6 Caring, leisure and other service occupations	-0.17	-0.02	-0.28
SOC: 7 Sales and customer service occupations	-0.24	-0.23	-0.16
SOC: 8 Process, plant and machine operatives	-0.53*	-0.01	0.43
SOC: 9 Elementary occupations	-0.30	0.27	0.27
Look-up answer selected: 2nd answer		-0.30*	-0.23
Look-up answer selected: 3rd-5th answer		-0.63***	-0.29
Look-up answer selected: 6th-12th answer		-0.92***	-0.33
Look-up answer selected: 13th-35th answer		-1.22***	-0.74***
Suitability of look-up code: Very well		0.75***	0.20
Suitability of look-up code: Not very/at all well		-0.90***	-0.26
Sex: Female	-0.14	0.14	-0.18
Highest qualification: Level 1 - GCSE lower grades or equivalent	-0.02	0.24	0.11
Highest qualification: Level 2 - GCSE higher grades or	-0.04	0.37**	0.33
Highest qualification: Level 3 - A-level or equivalent	-0.06	-0.11	-0.03
Highest qualification: Level 5 - Postgraduate degree	-0.10	0.19*	0.04
Highest qualification: Other academic qualifications	0.03	0.10	-0.09

Region: North East England	-0.13	0.27	0.07
Region: North West England	-0.11	0.02	-0.10
Region: Yorkshire and the Humber	0.04	0.29	-0.31
Region: East Midlands	0.00	-0.03	0.10
Region: West Midlands	0.10	0.09	0.09
Region: East of England	-0.03	0.10	-0.32
Region: South East	-0.13	0.18	0.03
Region: South West	-0.07	0.21	-0.10
Region: Other UK states	1.28*	0.49	-0.08
Index of Multiple Deprivation: 1st decile	-0.00	-0.13	0.05
Index of Multiple Deprivation: 2nd decile	-0.04	-0.12	-0.10
Index of Multiple Deprivation: 3rd decile	-0.15	0.07	0.03
Index of Multiple Deprivation: 4th decile	-0.03	0.18	-0.08
Index of Multiple Deprivation: 6th decile	-0.08	0.16	0.04
Index of Multiple Deprivation: 7th decile	-0.31	0.21	-0.15
Index of Multiple Deprivation: 8th decile	-0.16	-0.04	-0.23
Index of Multiple Deprivation: 9th decile	-0.22	-0.18	0.05
Index of Multiple Deprivation: 10th decile	0.11	0.10	0.03
Household size: 1 person	0.19	-0.09	0.08
Household size: 3 persons	0.07	-0.05	-0.20
Household size: 4 persons	-0.05	0.14	-0.05
Household size: 5 or more persons	0.12	0.07	-0.02
Housing tenure: Own outright	0.25	0.07	-0.33
Housing tenure: Rent it	0.02	0.00	0.13
Housing tenure: Live rent free	-0.08	-0.22	-0.00
Housing tenure: Other	-0.19	0.06	-0.27
Marital status: Married or a civil partner	0.20*	0.02	0.10
Marital status: Divorced	0.53	0.20	-0.06
Marital status: Other	-0.50	-0.42	0.16
Ethnicity: Mixed	-0.06	0.16	0.05
Ethnicity: South Asian	-0.30*	-0.11	-0.03
Ethnicity: Black	-0.11	0.09	-0.32
Ethnicity: Other	-0.22	0.05	-0.65
Constant	*2.45***	*0.89***	*3.08***
Pseudo R-Square (McFadden R2)	*0.096*	*0.112*	*0.048*

****p<0.001, **p<0.01, *p<0.05; Reference categories: Mode & device: Web, PC/Laptop, Length of job title: 1-13 characters, Length of keywords describing job: 1-19 characters, Length of open-descriptions of jobs: 1-38 characters, SOC: 2 Professional occupations, Look-up answer selected: 1st answer, Suitability of look-up code: Fairly well, Highest qualification: Level 4 - Graduate degree, Region: London, Index of Multiple Deprivation: 5th decile, Household size: 2 persons, Housing tenure: Own, buying with help of mortgage/loan, Marital status: Single, Ethnicity: White*



University of Essex

University of Southampton

UKRI Economic and Social Research Council

NCRM National Centre for Research Methods

Office for National Statistics

UCL

Warwick The University of Warwick

Manchester 1824 The University of Manchester

City University of London Est 1894

National Centre for Social Research

LSE The London School of Economics and Political Science

Ulster University

Unil UNIL | Université de Lausanne

Ipsos

verian Formerly Kantar Public

www.surveyfutures.net